EXPLORING DRIVING FACTORS OF HIGHER PAID TAXI TRIPS USING ORIGIN-DESTINATION GPS DATA (CASE STUDY: GREEN TAXIS OF NEW YORK CITY)

P. Mojtabaee[1, *], M. Molavi [1], M. Taleai [1]

[1] Faculty of Geomatics, K. N. Toosi University of Technology, Tehran, Iran. (pooya.mojtabaee@email.kntu.ac.ir; taleai@kntu.ac.ir; moein.molavi@email.kntu.ac.ir)

Commission VI, WG VI/4

KEY WORDS: GPS Trip Data, NYC Green Taxi, Spatial Patterns, Cost and Fare, Driving Factors, Socioeconomic

ABSTRACT:

Investigating the influential factors of the areas where people use taxis is a crucial step in understanding the taxi demand dynamics. In this study, we intend to analyze higher-paying taxi trips by putting forward an approach to explore a dataset of green taxi trips in New York City in January 2015 together with some demographic, housing, social and economic data. The final goal is to find out whether the chosen factors are statistically significant to be considered as potential driving forces of demand location for trips with a higher-paid fare. Since airports are major attracting sources for taxi travels, all the steps are taken separately for three scenarios that the trip drop-offs are in 1) LaGuardia Airport, 2) John F Kennedy Airport or 3) other areas. First, the spatial pick-up distribution of these higher-paying trips is mapped to enable visual comparison of the urban movement patterns. Then, taking into account the pick-up density as the response variable, the densities of: foreign-born's population, number of houses with no vehicles, the private wage and salary workers' population, the government workers' population and the self-employed workers' population in own not incorporate business were considered as the explanatory variables. These variables were examined to find important factors affecting the demand in each neighborhood and different results in each of the three scenarios were discussed. This study gives a better insight into discovering driving factors of higher-paid taxi trips when considering airports as destinations which attract travels with potentially different characteristics.

1. INTRODUCTION

Recent technological advancements in sensors and GPS devices have created an unprecedented opportunity to capture and store huge volumes of data from various sources. With the rapid growth of the urban population, it is more crucial to utilize this nowadays more available data in order to understand the social dynamics and urban behaviors. The data stored in the transportation sector can be very beneficial to this purpose. Taxi fleet plays an important role in urban transportation as they move freely across the city and can supply the demand wherever there is a need. Therefore, the data of GPS-equipped taxis can give us a clearer point of view on the mobility of people to help in decision-makers have a more detailed and more precise view in the area of city planning. These decisions could be applied in traffic management, provision of new infrastructure or to provide both taxi drivers and citizens with information on where to pick up passengers or find taxis.

New York City (NYC) comprises five boroughs of Manhattan, Brooklyn, Queens, Bronx and Staten Island. In this city, taxi fleets operate under regulations of the Taxi and Limousine Commission. Two important kinds of taxis in NYC are yellow and green taxis. Yellow taxis are the first and main part of the fleet allowed to pick passengers up and drop them off all over the NYC but as the yellow taxi drivers tended to pick up passengers in Manhattan below 96th Street and the two airports, the Taxi and Limousine Commission introduced green taxis to serve the demand and balance the supply coverage. These taxis are only allowed to pick people above East 96th Street and West 110th Street and in outer boroughs except the two airports unless the trips are pre-arranged.

There has been lots of research on taxi trip analysis from different points of view. In an analysis, Schaller (2005) used multiple linear regression models to predict the number of taxicabs in 118 United States cities and found some influential factors. In another paper, Mousavi et al. (2012) investigated trip generation for all modes of transit and stated some of the most influential factors such as the structure of the employment, household, age, income, gender, marital status, having a car, the density of population and the distance to transit. In a study, Austin and Zegras (2012) investigated if taxi fleet in Boston, Massachusetts complements the public transit system or works as an alternative to it. For the generation of taxi trips, they developed a Poisson count model by making use of taxi GPS data of a 4-day timespan and some demographic data of Boston.

Since transit and taxi fleet supply the need of the public, people's choice to use either of these modes were related to some factors in some studies. In their paper, Recca and Ratledge (2004) examined a list of potential factors on mode choice and concluded that in the city of Wilmington, Delaware, locations with high densities of employment and population have absorbed focus of high transit service. In a study by Corpuz (2007), the driving factors influencing citizen's modal choice between private vehicles and public transportation were investigated. It was stated that socioeconomic characteristics and the time of day were of high importance and that workers and households with better incomes tended to use private cars instead of public transit. It was also mentioned that based on the results, because of avoiding the traffic, bus and train are more

* Corresponding author

popular during the morning and late afternoon. Yang and Gonzales (2014) developed two different methods to model taxi pick-up and drop-off by the time of day in NYC and demonstrated its temporal and spatial variations in census tracts and proposed a technique by which they measured accessibility to discover the relationship between taxi demand and transit service. Another paper characterized people's travel movements such as travel distance, speed, time and the distribution patterns of origins and destinations on weekdays and weekends by using GPS data of 1100 taxis. (Yao and Lin, 2016). Alfayez and Aldawood (2017) presented an approach of land classification by relating taxi cost to TAZs. They analyzed their data based on gender and they explored their dataset visually. In a study by Yang and Gonzales (2017 they concluded that negative binomial distribution can better demonstrate the effects of their chosen explanatory variables on the taxicab pick-ups. Shen and Chen (2017) investigated people's spatiotemporal behavior patterns in Nanjing and modeled the relationship between the density of pick-up and drop-off locations and population density, transportation density and per capita disposable income. Another paper scrutinized the correlation between taxi demand, land use pattern and accessibility to other public transit modes in Washington, DC (Yang et al., 2018). However, having in mind that:

- NYC green taxi coverage and supply are formed and limited by the regulations as opposed to yellow cabs which are free to pick up their passengers everywhere in NYC,
- Discovering influential factors on higher-paying taxi trips has been poorly investigated,
- Considering airport trips as separate major kind of trips which the driving factors of the trips related to them have been of less attention in the literature,

this study aims to address and fill in these gaps. The remainder of this article is structured as follows. In Sect. 2 the steps and methodology are presented. The results are included in Sect. 3 and the conclusion is drawn in Sect. 4.

2. METHODOLOGY

2.1 Dataset

The taxi dataset used in this paper includes trip logs from the green taxi trips completed in NYC in Jan 2015 which is publicly accessible on the NYC Taxi & Limousine Commission website. Many attributes such as pick-up and drop-off geographical coordinates as well as time, distance, fare amount and some others have been recorded in this dataset.

Along with the taxi data, we used NYC shapefile at the Census Tract level (2165 census tracts) available at the NYC Opendata website and also American Community Survey data of 5-year estimate of 2016 for each census tract in housing, social and economic categories published at the United States Census Bureau website.

2.2 Data Pre-processing

The total number of trips in the dataset is 1508501. A pre-processing step has been performed to remove invalid data such as trips with geographical coordinates which are out of the NYC boundary, trips with the same pick-up and drop-off coordinates, trips with zero or negative travel distance, trips recorded with no passenger, trips with zero or negative fare amount and trips having unusual travel distance comparing to their trip duration time. For the latter, the criterion of the maximum trip average speed of 100 miles/hr was considered and all the trips having the average speed above this limit were excluded. The total number of trips remaining after pre-processing is 1474166 and therefore the invalid data comprises about 2.3% of the initial data.

2.3 Defining Higher Paying Taxi Trips

After cleaning the dataset of invalid data, a limit value is set in order to separate the trips with higher payments. Median for the fare amount is chosen as the threshold value and the reason is that the median is resistant to outliers and it is not affected by extreme values with low frequency comparing to the mean. The median fare amount in the dataset is calculated and it is equal to 9 dollars. Hence, the trips above this value are extracted regarded as the higher paying trips for the next steps and they become the target dataset.

2.4 Spatial Distribution of Higher Paying Trips

In this step, the total number of pick-ups located in each census tract is obtained. Then the density of these pick-ups in these polygons is calculated. With the notion that the trips with airport drop-off comprise a considerable portion of the taxi trips and that they might demonstrate different patterns from those ending in different destinations, the spatial distribution of taxi pick-ups for each of these destination cases are shown in separate maps. The classification method to visualize the data which is utilized here is the quantile method and the first class is considered as the least significant and is colored as grey so that the other classes could be visualized in a better way (Fig. 1&2&3).

2.5 Demand Model

In this study in order to discover the relationship between the explanatory variables and the response variable, a multiple regression model is employed (Eq.1). Multiple regression is generally used to predict a dependent variable by several independent variables or to explain the relationship between them. The multiple regression equation takes the following form:

$$y = \beta_0 + \beta_1 x + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon \qquad j = 0,1,\cdots,k \tag{1}$$

where β_j are unknown parameters which represent the expected change in the response y per unit change in x_j when all other predictor variables are held unchanged. These β_j are called regression coefficients, x_j are regressors or predictor variables, β_0 is called the intercept and ε is the error accounting for the failure of the model to exactly fit the data.

R-squared or the coefficient of determination is a measure for the goodness-of-fit. In other words, it evaluates how close the data are to the fitted regression line and shows the percentage of the variance in the dependent variable that can be predicted by the independent variables.

In multiple regression, multi-collinearity happens when two or more explanatory variables are highly linearly related and therefore redundant. One of the popular ways to measure multi-collinearity is the Variance Inflation Factor (VIF) which assesses the multi-collinearity by measuring how much the variance of an estimated regression coefficient increases if your predictors are correlated. A VIF of 1 for a factor means that there is no correlation among that predictor and the remaining predictor variables. In case of multi-collinearity (having a VIF above 5 for a factor) the highly correlated predictor could be removed from the model.

Here, the pick-up density in each census tract representing the demand is considered as the response variable and the explanatory factors are the population density, density of private wage and salary workers, density of government workers, the density of self-employed workers in own not incorporate business and the density of unpaid family workers. These densities are calculated by dividing the population value of these data by the area of each census tract.

3. RESULTS

3.1 Visualization Results

Figure 1 and Figure 2 show the pick-up density for trip destinations of LaGuardia Airport and JFK Airport respectively, and Figure 3 depicts the spatial distribution of taxi pick-up densities for the trips ending at other locations.
The results of these three different cases show that visually the spatial distribution of pick-up density for the two airports are almost the same but the spatial distribution of pick-up density for other destinations covers a larger area although it still indicates relatively the same major pick-up areas as the other two scenarios.

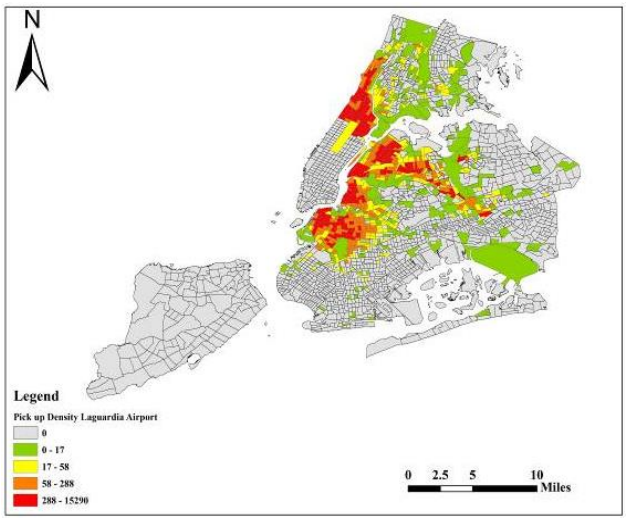

Figure 1. Pick-up Density for LaGuardia Airport

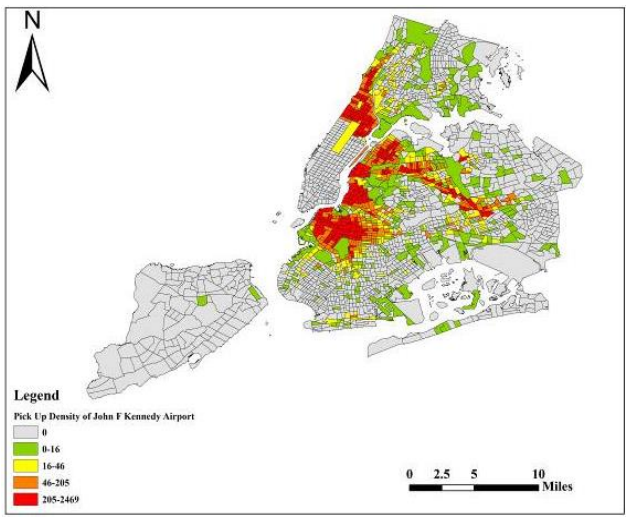

Figure 2. Pick-up Density for JFK Airport

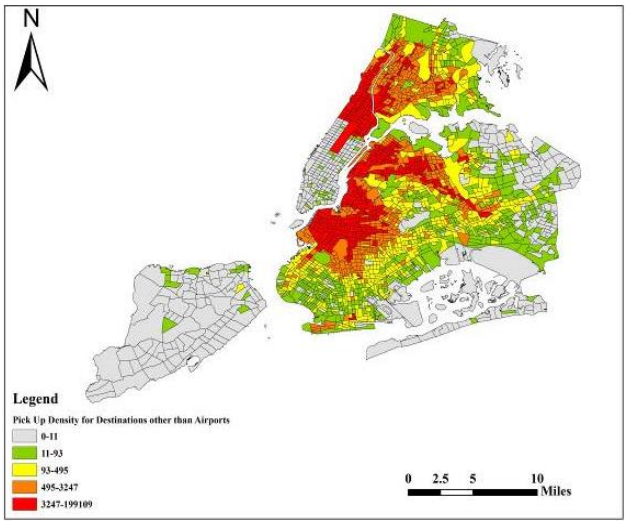

Figure 3. Pick-up Density for Destinations other than Airports

3.2 Regression Results

Three multiple linear regression models are formed for the three cases discussed above to discover the relationships between the variables. It is important to mention that some census tracts were related to cemeteries, island, parks, etc. and the data of the considered factors were not available for those tracts, hence they were eliminated from the calculations. Moreover, the data used in this paper is only a one-month log of green taxis, and therefore many census tracts lacked any number of pick-ups, or they had very few ones, so the census tracts having less than five number of pick-ups were also excluded from entering the model. The respective results for the three multilinear regression models are presented in Table 1, 2 & 3.

Term	Coef.	P-Value	VIF
Constant	414	0.0004	
Density of Houses with No Vehicle	0.0139	0.0818	2.8
Density of Government Workers	-0.0849	0.0157	2.39
Density of Self Employed Workers in Own not Incorporate Business	0.0747	0.0503	1.51

Table 1. Regression Results for LaGuardia Airport Pick-up Location Driving Forces

The R^2 value for LaGuardia Airport model was 3.05% and the p-value results suggest that the density of government workers and the density of self-employed workers in own not incorporate business are showing a significant relationship with our response variable. But the density of houses with no vehicles is above the significance level of 95% so it can show no real relationship with pick-up densities in each census tract. For the second case which is JFK airport, the R^2 value is 16.95% and except the density of unpaid family workers, all the factors presented in Table 2 have a statistically significant effect on the dependent variable at the chosen level. Finally, the last regression model which is for the trips ending at areas other than the two airports, the R^2 of 20.37% is obtained and all the four variables presented in Table 3 show statistically strong relationship on the taxi demand.

Term	Coef.	P-Value	VIF
Constant	261.1	0	
Houses with No Vehicle Density	0.01545	0.0000000000097245	3.3
Foreign Born Density	-0.00467	0.0000071126346915	1.94
Government Workers Density	-0.04793	0.0000001575468365	2.31
Self Employed Workers in Own not Incorporate Business Density	0.0303	0.002691063	1.7
Unpaid Family Workers Density	0.092	0.436880887	1.05

Table 2. Regression Results for JFK Airport Pick-up Location Driving Forces

Term	Coef.	P-Value	VIF
Constant	4579	0.0000000015524394	
Houses with No Vehicle Density	0.7433	0	2.98
Foreign Born Density	-0.3566	0	1.99
Government Workers Density	-1.406	0.0000000093719005	1.84
Self Employed Workers in Own not Incorporate Business Density	3.356	0	1.8

Table 3. Regression Results of Pick-up Location Driving Forces for Trips with Destinations other than Airports

Achieving the R^2 values of 3-20% is not unusual since modeling taxi trips with high predictability is a complicated problem demanding consideration of many factors and the goal of this study was not to estimate the dependent variable precisely but to find potential new driving factors.

All the factors mentioned in subsection 2.5. were entered in the initial model but some of them such as population density and the density of private wage and salary workers were removed and not mentioned in the final results because of having large VIF and therefore indicating multicollinearity with other variables.

4. CONCLUSION

In this study, a one month logged green taxi trips of Jan 2015 along with some social, economic and housing data at the census tract level were used to explore the trips which are having higher payments at their destinations and to find some influential factors on demand for these trips. These steps were taken based on the idea that due to the importance of JFK and LaGuardia airports as significant sources attracting taxi trips, these trips might possess different characteristics. Although pick-up density maps of these two airports do not show a substantial difference, the relatively high difference in their regression model results suggests that if the aim of the research is to reach high predictability of a model for taxi demand, they might need different models of their own. In this paper, some influential factors were identified and some differences in the three scenarios were revealed. For future research, these steps can be combined with the yellow taxi trips which are free in their coverage, taxi trip data of longer timespan, land use data, transportation layers, and consideration of temporal factors to yield more information and better prediction ability.

REFERENCES

Alfayez, A. and Aldawood, S., 2017, July. Visual Exploration of Urban Data: A Study of Riyadh Taxi Data. In International Conference on Social Computing and Social Media (pp. 327-337). Springer, Cham.

Austin, D. and Zegras, P.C., 2012. Taxicabs as public transportation in Boston, Massachusetts. Transportation Research Record, 2277(1), pp.65-74.

Corpuz, G., 2007, September. Public transport or private vehicle: factors that impact on mode choice. In 30th Australasian Transport Research Forum.

Mousavi, A., Bunker, J.M. and Lee, B., 2012. A new approach for trip generation estimation for use in traffic impact assessments. In 25th ARRB Conference Proceedings. ARRB Group Ltd.

Ratledge, E.C. and Racca, D.P., 2004. Project Report for" Factors That Affect and/or Can Alter Mode Choice".

Schaller, B., 2005. A regression model of the number of taxicabs in US cities. Journal of Public Transportation, 8(5), p.4.

Shen, J., Liu, X. and Chen, M., 2017. Discovering spatial and temporal patterns from taxi-based Floating Car Data: a case study from Nanjing. GIScience & Remote Sensing, 54(5), pp.617-638.

Yang, C. and Gonzales, E.J., 2014. Modeling taxi trip demand by time of day in New York City. Transportation Research Record, 2429(1), pp.110-120.

Yang, C. and Gonzales, E.J., 2017. Modeling taxi demand and supply in New York City using large-scale taxi GPS data. In Seeing cities through big data (pp. 405-425). Springer, Cham.

Yang, Z., Franz, M.L., Zhu, S., Mahmoudi, J., Nasri, A. and Zhang, L., 2018. Analysis of Washington, DC taxi demand using GPS and land-use data. Journal of Transport Geography, 66, pp.35-44.

Yao, C.Z. and Lin, J.N., 2016. A study of human mobility behavior dynamics: A perspective of a single vehicle with taxi. Transportation Research Part A: Policy and Practice, 87, pp.51-58.